BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

An In-Band Power-Saving Protocol for Mobile Data Networks

Apostolis K. Salkintzis and Christodoulos Chamzas, *Senior Member, IEEE*

***Abstract*—A new alternative is proposed for reducing the power consumption of the portable (battery-powered) units operating in a mobile packet-data network. First, a short review of the current power-saving protocols is taking place. It is shown that the most common means for conserving power is the intermittent operation of the receivers (at the portable units) and a central administration authority that synchronizes the receivers. Some drawbacks of the synchronous operation lead us to the introduction of an asynchronous power-saving protocol, where no central synchronization is necessary and where each terminal may control its power consumption relative to its current needs. According to the proposed power-saving page-and-answer protocol, an acknowledgment paging procedure is preceding every packet transmission in order to alert mobile terminals with pending traffic. Steady-state performance is evaluated with the aid of simulation. The relationship between the achieved power-saving and the mean packet delay degradation is presented. Finally, we express some notable implementation issues and some considerations regarding the employment of this protocol as a supplementary power-saving service in microcellular mobile data networks and wireless local area networks.**

I. Introduction

THE PRINCIPAL objective of a power-saving protocol is to minimize (by software means) the power consumption of the "power-sensitive" (e.g., battery-operated) network elements and, *concurrently*, to keep the performance trade to a minimum.

The main idea behind software-controllable power consumption is the discontinuous reception, that is, the ability of terminals to periodically power down their receivers. The way discontinuous reception may be implemented depends highly on system peculiarities, so different approaches are usually adopted by different systems. While discontinuous reception is an easy concept, it raises a number of problems and special means should be provided to cope with them. For example, in a mobile packet data network discontinuous reception causes mobile terminals to periodically become unavailable for reception; thus, any communication with them should be prohibited during these "deaf" periods. Therefore, discontinuous reception causes an unavoidable performance (mainly throughput) degradation. The main engineering concept in this case is the optimum throughput–power-saving balance.

Discontinuous reception has been extensively utilized in a number of dissimilar systems ranging from paging systems to cellular and cordless systems and to mobile data networks. However, in each case a different implementation is constructed and different support means are provided. Let us make a short review.

In the old 5/6-tone paging system [1] an alert mechanism is used to wake up a group of pagers (each pager has a hardcoded group ID). Every group service period is preceded by a transmission of a group-specific tone that alerts all of the units belonging to the group which is about to be serviced. Units that do not belong to the currently serviced group can remain in standby. The main drawback of this scheme is the large overhead of the alert tone, specifically when little traffic is distributed to many groups.

In the digital POCSAG code [1] a suitable batching scheme is employed to increase the code's power-saving efficiency. In this system, batches are transmitted one after the other as long as there is pending traffic and each portable unit operates its receiver only during a specific (precoded) frame inside each batch. In other words, every unit is synchronized and wakes up only at specific time instants, where its messages are expected. After waking up, a portable remains in operation provided that a message for it is detected. All the others return to power-saving mode. In this case a considerable power saving is gained and the overhead, compared to 5/6 code, is significantly reduced. However, throughput remains small mainly because of the code's inflexibility.

In all of these paging schemes the system throughput is greatly degraded due to power-saving procedures, but this is quite acceptable under medium loading conditions and under the attractive gain of battery life extension. However, most modern paging systems [4] (like Motorola's FLEX [1] and the pan-European standard ERMES) employ more sophisticated codes and obtain a considerable improvement in the throughput–power-saving balance.

In mobile data networks (that we are mainly interested in), power-saving is based on some sort of synchronization. The usual approach is to have a base station periodically transmit a pending traffic list and ensure that the wakeup instants of mobile terminals are synchronized with these transmissions.

In MOBITEX [2], [3], [6] network, for example, special link frames ((SVP6) frames) are periodically transmitted by

Paper approved by R. A. Valenzuela, the Editor for Transmission Systems of the IEEE Communications Society. Manuscript received November 6, 1995; revised July 22, 1996 and May 31, 1997. This work was supported in part by GSRT under the PENED95/687 Program. This paper was presented in part at COMCOM5, Rethymnon Crete, Greece, June 26–30, 1995.

A. K. Salkintzis is with the Electrical and Computer Engineering Department, Democritus University of Thrace, Xanthi 67100 Greece (e-mail: salki@ee.duth.gr).

C. Chamzas is with the Image Processing and Multimedia Laboratory, Electrical and Computer Engineering Department, Democritus University of Thrace, Xanthi 67100 Greece (e-mail: chamzas@ee.duth.gr).

Publisher Item Identifier S 0090-6778(98)06643-8.

the base station in order to inform the portable fleet about the downlink traffic demand. Each (SVP6) frame contains a downlink traffic list, commonly composed of the addresses of all portables that have pending data packets in the base station. All portables are synchronized with these (SVP6) frame transmissions and wake up right before the transmission starts. When a portable identifies that there is buffered traffic for it, it remains in operation so the base station may forward all the buffered downlink traffic after the transmission of a (SVP6) frame.

Similar mechanisms are provided in CDPD [12] and IEEE 802.11 [13], [14] standards. However, in these systems a terminal is not forced to wake up whenever the traffic list is announced. Rather, it may choose to skip some announcements in order to further reduce its power consumption. In this case, though, it is not implicitly known when a terminal will become ready for reception, as in MOBITEX. In order to cope with this problem a terminal that gets ready for reception transmits a specific link frame to notify the base station that it can accept its buffered traffic.

Considering all of the above we can express the following observation—the power-saving protocols commonly employed today are *centrally controlled* protocols that offer power-saving features by means of synchronization. Wakeup instances are under *central administration*; therefore, every portable unit cannot *independently* set its power consumption level and it must obey the rules set by the network.

This approach of centrally controlled power saving has both advantages and disadvantages. The main advantage is the better utilization of system resources, and that makes it attractive to network operators. By having the base station controlling whenever the portable receivers will switch on and off, it is not likely to have overcrowded queues and, generally, the downlink processes can be tuned up easily. On the other hand, central control causes power saving to be equally distributed to all units (except for CDPD and IEEE 802.11) and this might not be the most favorable scheme for the users. Also, whenever some users have high pending traffic and the base station decides to transmit more frequently the traffic list message (to offset the queue development), the low traffic users will pay an extra power consumption (because they will wake up more frequently), so some unfairness will result. Furthermore, in the case of CDPD and IEEE 802.11, where terminals transmit notifications to inform the base station that they are ready to receive, instances with high uplink (from terminals to base) traffic demand may be generated. This high demand may be produced right after the transmission of a pending traffic list, if many terminals try to simultaneously notify the base station of their reception availability.

Ideally (at least for users), an alternative scheme could be employed where each user would have the capability to tune its power consumption level according to its current needs (its battery state for example). That would be most suitable for the users because it could offer extended system utilization even during low battery rating periods. In order to implement this scheme we need an asynchronous environment. There is no point in trying to synchronize the users because they need the freedom to implement their own standby-operation cycle. Additionally, an asynchronous scheme is easier to implement since timekeeping and accuracy requirements needed for synchronization are totally unnecessary.

The purpose of this paper is to study the performance of an asynchronous distributed power-saving protocol, namely, the in-band page-and-answer protocol. The main objective of this protocol is to allow mobile terminals to implement their own independent and dynamic power management algorithm[1] and obtain an optimum balance between packet delay and battery consumption. In such cases a terminal will be able to select a proper duty cycle for its receiver. For example, at healthy battery periods, a long duty cycle (or even a continuous reception) could be selected to maximize downlink throughput, whereas at low battery conditions, a short duty cycle could be selected to offer an increased service utilization even with degraded performance.

The rest of the paper is organized as follows. In Section II we provide a verbal description of the in-band power-saving page-and-answer (PSPA) protocol. In Section III some modeling details and assumptions that have been used in our simulation are presented. In Section IV we evaluate the performance of the protocol by illustrating some of its properties, like the mean packet delay, the packet delay variance, and the power consumption. We also consider the performance under two service disciplines, namely, the exhaustive and the nonexhaustive service discipline. Finally, in Section V we express our conclusions.

II. Description of the In-Band Power-Saving Protocol

Among the various page-and-answer protocols that we have considered [15]–[18], we confine ourselves here to the in-band case.

The in-band PSPA protocol describes a downlink operation mode that aids mobile terminals to achieve power saving. It may be viewed as an intermediate operation mode between synchronous TDM and asynchronous packet switching. The first arrangement (TDM) inherently provides some means for power saving (since terminals accept data only at predefined time slots) but it is also characterized by poor resource utilization and poor throughput [7] under bursty or low traffic. The second arrangement (common packet switching) exhibits significantly improved performance [9], [10] but receivers must operate continuously and therefore no power-saving features are provided.

We assume a simple network topology where our system operates in a wireless centralized environment, as the one depicted in Fig. 1, with separate uplink and downlink channels. Each terminal (mobile or portable) features limited power resources and it is considered to roam[2] around a base station with which it can establish a two-way communication link. Terminals are free to operate their receivers whenever they like to; they are not synchronized to a common reference clock and, therefore, they do not necessarily wake up at the same time (except if they happens to). Under this anarchy,

[1] However, this algorithm is outside the scope of this paper.

[2] Roaming or even mobility should not necessarily hold true. Our main consideration is the limited power resources but since these are typical to mobile networks, a power-saving protocol can be of real value to such systems.

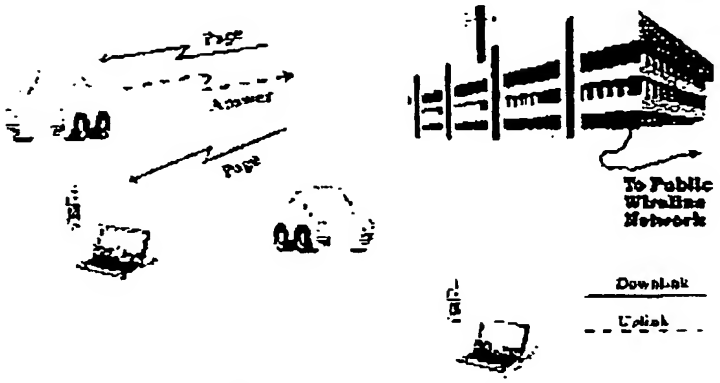

Fig. 1. Topological view of the considered mobile network.

a page-and-answer procedure is implemented to coordinate the communication with the base station. The procedure is as follows.

Packets destined to terminals with unknown reception state are temporarily buffered in the base station. Whenever there is any buffered traffic, the base station transmits paging messages that identify the downlink traffic demand: in other words, they identify the terminals that the buffered traffic should be forwarded to. The paging performed by the base station is continuous. One paging message succeeds the other until one or more terminals declare that they are ready for reception. In this case paging is suspended and the buffered traffic for these ready terminals is transmitted. After data transmission, paging resumes provided that there is residual buffered traffic[3].

A terminal willing to check if there is pending traffic for it powers up its receiver and monitors the downlink channel for some time. In case it receives a paging message including its own address it remains in receive mode and it also notifies the base station that it is available for reception. This is done by the transmission of a short notification message which we call an "acknowledgment." In this way the base station learns its reception availability and proceeds to data transmission. Of course, when the terminal receives no information indicating the presence of pending traffic, it switches back to power-saving mode.

As already mentioned, every paging message encodes specific information to aid remote terminals to identify which has pending packets in the downlink queue and, thus, which should wake up and prepare for data reception. We may visualize a paging message as a binary bit map that contains ones to the positions that correspond to terminals with pending traffic. A similar encoding scheme is used in the IEEE 802.11 standard [13], [14].

In the simplest case a paging message may encompass only a sequence of terminal addresses, while in other cases some complimentary data, such as priority tips, could be appended. Paging messages are supposed to be quite small and less costly than real data packets. Also, in the in-band arrangement (as opposed to the out-band arrangement [15], [18]) they share the same communication means with the data packets, so the downlink channel alternates between paging and data transmission periods. Depending on the encoding scheme, the length of the paging message can be considered either fixed or demand-proportional (e.g., growing with the number of the users that should be alerted). In this work, however, only the first case is assumed (*fixed-length paging messages*), for reasons that will come apparent in Section III.

Now consider what happens with the acknowledgments sent by the mobile terminals to declare their reception availability. Whenever a terminal discovers any pending traffic for it, a receiver-ready acknowledgment message is prepared for transmission. Yet, the acknowledgment does not generally supply an instantaneous feedback because it might need to compete for channel resources or it might be corrupted by the wireless channel impairments. However, we assume here[4] that all of the acknowledgments arrive at the base station uncorrupted and also arrive right after the end of a paging message transmission (i.e., without delay).

Fig. 2 represents a snapshot of the in-band PSPA protocol and visualizes all of the aforementioned statements. The first three waveforms represent receiver's activity (high level indicates a receiver in operation; low level indicates a receiver in sleep mode) in three typical terminals A1, A2, A3, while the bottom line represents the activity of the downlink channel. Arrows at the bottom represent packet arrivals.

According to this figure, when the first arrival (for A1) occurs, a paging sequence starts transmission. Note that, when a packet for A2 arrives, paging messages are updated to indicate pending traffic for A2 too. The fifth paging message is being captured by A2 and its acknowledgment arrives immediately at the base station. Afterwards, a data service period for A2 takes place.

In cases where two or more acknowledgments arrive simultaneously at the base station (like from A1, A3, as shown in

[3] Of course, in a real system the base station won't page a given terminal forever. After a maximum paging period, it will assume that the terminal is shut down and it will delete its ID from the paging messages.

[4] An accompanying paper will address protocol performance without these assumptions.

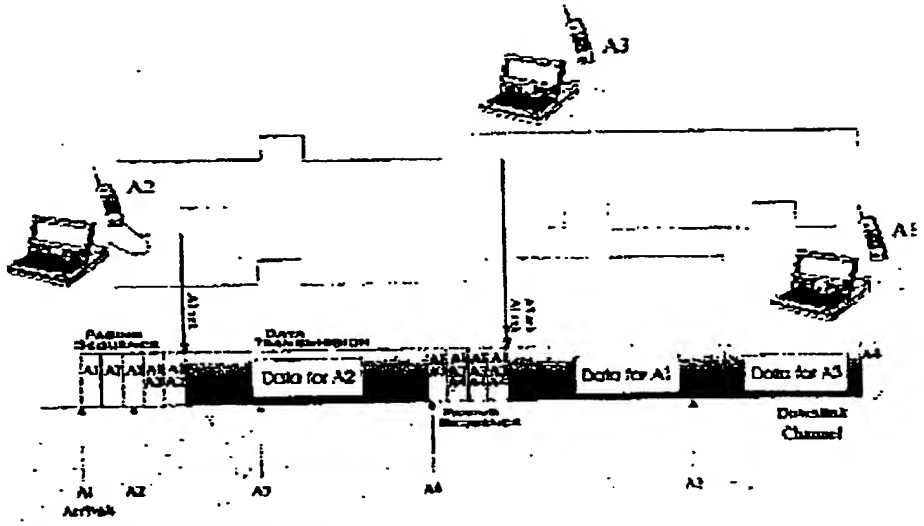

Fig. 2. Timing diagram for the in-band PSPA protocol.

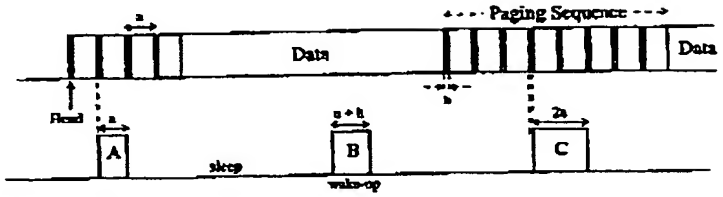

Fig. 3. The variation of wakeup period.

Fig. 2) the ready terminals are served either with priority or random order. In Fig. 2 for example, A1 is serviced first and A3 second.

As indicated by Fig. 2, when a terminal wakes up during a data transmission, it keeps its receiver up for a short period and then goes back to sleep mode. Therefore, when the system is heavily loaded, it is possible for a terminal to always miss a paging sequence. This means that the protocol can not guarantee an upper bound to the packet delay and that the packet delay variance at high traffic rates is expected to be considerably increased.

III. The Simulation Model

Our model consists of N identical mobile terminals. Packets for these terminals arrive at the base station according to a Poisson distribution with a *total* mean rate $\bar{\epsilon}$. Each packet has an exponentially distributed length with mean value L and we select our time unit to be the transmission time of the mean data packet. All of the arriving packets are queued in a buffer with infinite capacity and the base station transmits short paging messages to alert the terminals that have pending packets in this buffer. When one or more terminals declare that they are ready for reception, the base station service them either exhaustively or nonexhaustively, as we will explain later.

When a terminal is not in receive mode (has not sent an acknowledgment), it performs a wakeup–sleep cycle with a duty cycle $q = t_w/(t_w + t_s)$, where t_w is the *average* duration of the wakeup period and t_s is the duration of the sleep period (which is constant). Remember that a terminal goes back to sleep mode after receiving a paging message that indicates it has no pending traffic. So, generally, the wakeup period will not be constant. In the best case (see case A in Fig. 3) it will be as small as the duration of a paging message α, and in the worst case (case C in Fig. 3) it will be approximately equal to 2α.

Moreover, if a paging message is not detected during a time period t_{wm}, a terminal switches back to sleep mode (see case B in Fig. 3). We select the duration t_{wm} to be sufficiently long[5] in order to *guarantee* the reception of one paging message, when the wakeup occurs during paging transmission. This will minimize the number of lost paging messages and, consequently, the mean packet delay.

[5]The maximum time needed to detect a paging message is $t_{wm} = \alpha + h$, where h is the duration of paging message header.

Since t_{sm} depends on the duration of a paging message a, we choose a to be constant—otherwise a terminal would have to find out the length of paging messages in order to remain for long enough in a wakeup state. To avoid this complexity we assume that all paging messages have equal length, and that clarifies our previous statement of Section II.

Another assumption that we make in our simulation model is that the start of *every* new paging sequence is uniformly distributed in a terminal's cycle period. Although the wakeup–sleep cycle is periodic, this allegation holds true as long as the base station loses track of the terminals' timing between successive paging sequences. This is again realistic, since both the paging sequence duration and the data transmission duration are random.

In the results that we present below, all of the terminals are assumed to operate with the same duty cycle q unless otherwise stated. This is not very realistic in a system where terminals are free to choose their own sleep interval t_d, but it is sufficient for demonstrating the power-saving and packet delay characteristics of the protocol. A case with different duty cycles is also considered.

IV. Simulation Results

A. Bandwidth Allocation

Our system exhibits an important advantage when compared to other synchronous schemes (like CDPD and IEEE 802.11). Specifically, it does not produce high peaks to uplink demand since there is no wakeup synchronization, and it is unlikely that many terminals will concurrently try to transmit acknowledgments.

The lack of wakeup synchronization features another interesting property—the paging traffic becomes adapted to the total downlink demand. That is, when the downlink demand is increased (we page more and more terminals), paging traffic is decreased; therefore, paging does not consume excess communication resources when we need them. On the other hand, when the downlink demand is small, paging occurs for a considerable duration (because we page a small number of users) and uses considerable resources. However, this is harmless because the channel is mostly idle in this case.

This adapted behavior of the paging traffic is evident from Figs. 4 and 5, where the channel time (percentage) occupied by the paging messages is depicted as a function of the offered downlink traffic for various values of duty cycle q and paging message length a.

According to Figs. 4 and 5, it is clear that the necessary communication resources needed to support the paging process are not the same for every offered traffic. As the transmission demand increases, more paging messages are delivered and more channel time is consumed. However, at a critical traffic point, where many users are being paged simultaneously, the paging sequences start to reduce in length because acknowledgments arrive more often. For an exhaustive service (explained below), paging tends to vanish as we approximate high traffic rates.

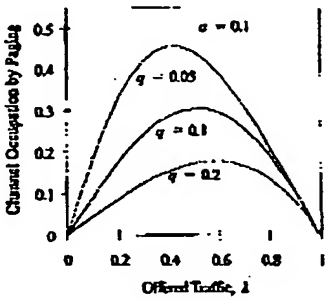

Fig. 4. Channel occupation by paging for various values of receiver duty cycle q.

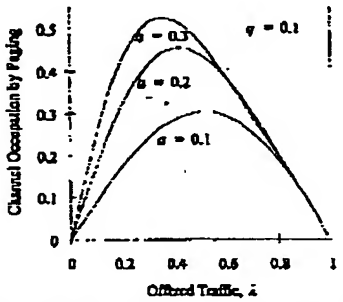

Fig. 5. Channel occupation by paging for various values of paging message length a.

It is interesting to note that the curves shown in Figs. 4 and 5 are *independent* of the number of terminals N, provided that the value of a is constant (as we have assumed). That means that the resources required for paging depend only on q, a, and total traffic $\mathcal{E}$. Indeed, if N increases (but the total downlink demand remains the same), we will page more users at a time because the incoming packets will be distributed to more users; thus, the paging sequences will become smaller. On the other hand, we will need more paging sequences to support a given traffic, so combining the two effects, we end up with the same channel occupation distribution.

B. Service Discipline

We consider now two service disciplines, namely, the exhaustive and the nonexhaustive service disciplines, and we study each case separately.

1) Exhaustive Service: In the exhaustive service the base station services exhaustively a terminal that has sent an acknowledgment, i.e., it transmits *all* of its pending traffic in a burst. Moreover, if two or more terminals send acknowledgments simultaneously, all of them are exhaustively serviced before going back to paging. The order that the terminals are served may be either random or prioritized. This discipline is summarized in the flowchart of Fig. 6.

The main advantage of the exhaustive service is that the terminals remain in receive mode for a short time—as long as it is needed to receive all their buffered traffic. This is a

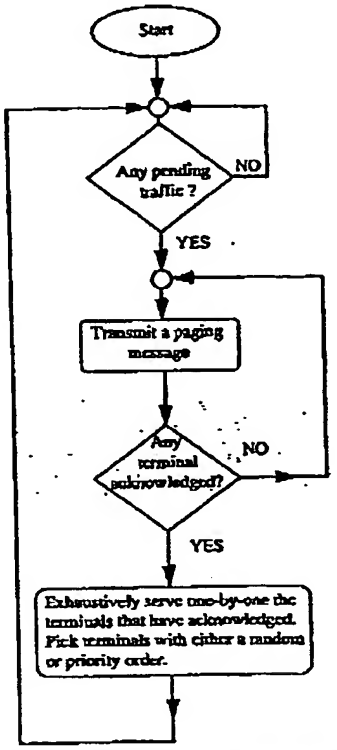

Fig. 6. Flowchart for the exhaustive service discipline.

beneficial property since we are interested in power saving. However, the exhaustive service inherently introduces some sort of unfairness—terminals with high buffered traffic may monopolize the downlink channel for long periods. For this reason, we also consider a nonexhaustive service discipline later on.

a) Packet delay: In Figs. 8 and 9 the mean packet delay of the PSPA with exhaustive service (PSPA/E) protocol is presented. Curves in Fig. 8 correspond to various paging message lengths[6], while curves in Fig. 9 correspond to various duty cycles. We also plot the mean packet delay of the M/M/1 system [9], [10] that corresponds to the common case, where all receivers are continuously in operation. In this way we may compare the downlink performance of the PSPA/E protocol against any packet data network that implements no power saving. It is important to note that the M/M/1 case that the results are compared to is an ideally performing case.

We have assumed no priority for the curves of Figs. 8 and 9. That is, whenever many terminals acknowledge the reception of the same paging message, the base station services all of them in *random* order.

[6] Actually, α is equal to the transmission time of a paging message. Remember that our time unit is equal to the transmission time of the mean data packet.

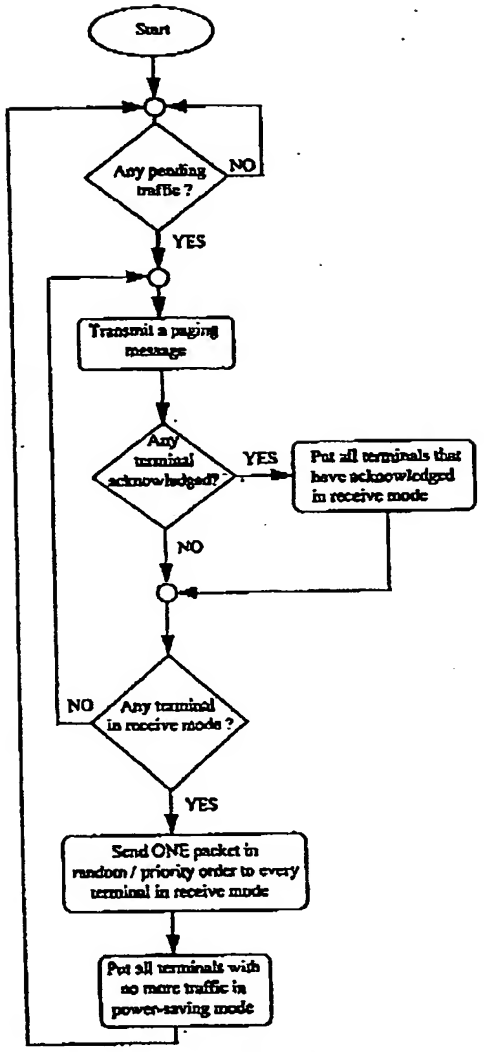

Fig. 7. Flowchart for the nonexhaustive service discipline.

Observe from Fig. 8 that the average message delay depends highly on the length of paging message. We should make this length as small as possible in order to meet a small mean packet delay and to enhance the system performance.

In Fig. 9 we show how the average packet delay is related to the receiver's duty cycle. We see that employing a 0.05 duty cycle (translated to a 95% power saving in the receive unit at low traffic rates) the mean packet delay is degraded by a factor of about 3.4 (when $a = 0.1$) over *all* of the practical traffic conditions. When the duty cycle is 0.1, the degradation factor (D_F) falls to two, and when the duty cycle is 0.2, D_F stays below 1.4. The careful reader may observe that the average packet delay does not degrade linearly with the duty

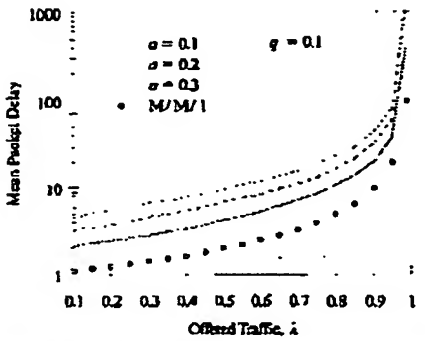

Fig. 8. Mean packet delay for the PSPA/E protocol under various paging message lengths.

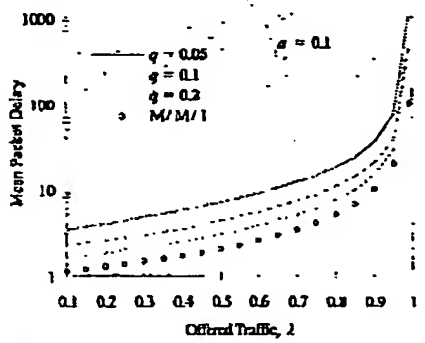

Fig. 9. Mean packet delay for the PSPA/E protocol under various duty cycle values.

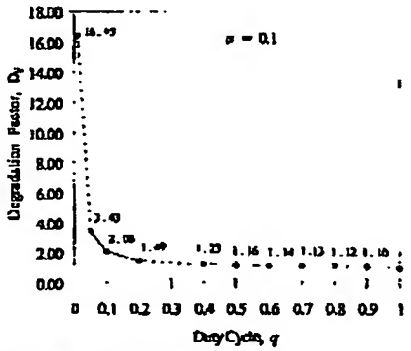

Fig. 10. Mean packet delay degradation factor.

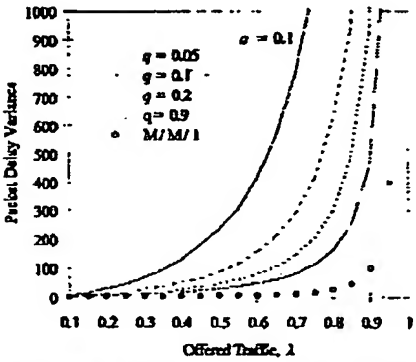

Fig. 11. Packet delay variance of the PA/E system for various duty cycles.

cycle. This is absolutely reasonable since the duty cycle is not linearly related to the receiver's sleep period (assuming a constant wakeup interval).

Fig. 10 displays how the degradation factor D_F (mean packet delay in the PSPA/E protocol to the mean packet delay in M/M/1 case, under the same loading) varies for full-range duty cycle coverage. It is interesting to note that there is an almost flat range where the variation of the duty cycle does not cause significant effects on the mean packet delay. In this range the receiver's sleep interval is not excessively long as to introduce large packet delays. However, for small duty cycles (say, smaller than $q = 0.1$), a significant increase of mean packet delay is developed because the sleep interval becomes long.

b) Variance: In Fig. 11 we focus on the variance of the packet delay. Specifically, the packet delay variance is illustrated versus total downlink traffic demand for four values of receiver's duty cycle, namely, 0.05, 0.1, 0.2, and 0.9.

The variance, as compared to the M/M/1 case (with a first-come first-served (FCFS) service discipline [10]), develops a considerable increase, especially when the duty cycle is small. This extra variance is introduced by the paging process—a terminal may send acknowledge either in a very short time—when the first paging message happens just before its wakeup instant—or after a considerably large time—when other terminals happen to acknowledge and be served before it. Thus, whenever many terminals are being paged, the distribution of the paging delay is expected to be wide or, alternatively, the variance of the packet delay is expected to be large. Observe that when the duty cycle is small, many terminals are being simultaneously paged even when the total downlink traffic is small. That is why the curves of Fig. 11 that correspond to small duty cycles ($q = 0.05$ or 0.1) show a significant variance increase.

More specifically, when $q = 0.05$ and $\bar{e} = 0.1$, the variance is as much as 3.2 times the corresponding variance of the M/M/1 system, while for $q = 0.1$ and 0.2 the increase is 1.5 and 1.1, respectively. Moving up to $\bar{e} = 0.3$, the increase for $q = 0.05$ is about 32.8 and for $q = 0.1$ and 0.2 is roundly 13.3 and 7.4, respectively. The same scenario applies for all the range of practical traffic demands.

A main issue, originating from these considerations, is that the PSPA/E protocol features a large packet delay variance when the receivers implement a small duty cycle. Effectively, this will make cumbersome the implementation of time-sensitive applications, and this is another cost that we should be prepared to pay for if we need a very large power saving. In such cases, applications should encompass extra intelligence to cope with the wide packet delay distribution

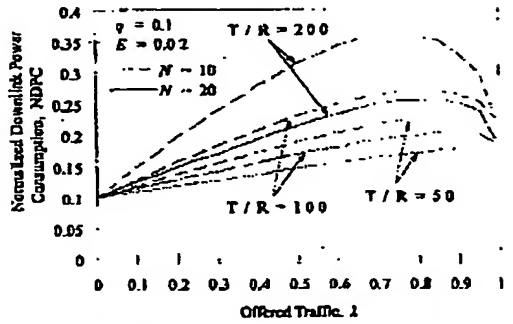

Fig. 12. NDPC for various transmit-to-receive power ratios.

TABLE I
AVERAGE NUMBER OF TERMINALS SERVED PER SERVICE PERIOD

Total offered traffic λ	Average number of terminals served per service period			
	$N = 10$		$N = 50$	
	$q = 0.1$	$q = 0.2$	$q = 0.1$	$q = 0.2$
0.1	1.0050	1.0051	1.0050	1.0056
0.2	1.0117	1.0139	1.0135	1.0152
0.3	1.0201	1.0239	1.0249	1.0308
0.4	1.0306	1.0411	1.0407	1.0522
0.5	1.0452	1.0621	1.0640	1.0847
0.6	1.0637	1.0916	1.1007	1.1348
0.7	1.0901	1.1334	1.1550	1.2243
0.8	1.1252	1.2017	1.2616	1.3773
0.9	1.185	1.5244	1.5127	1.7788
0.95	1.3761	1.5373	2.6637	3.8421

and, ideally, new automatic-repeat-request (ARQ) algorithms should be implemented to optimize link-level performance. Various priority policies could be also implemented to make the use of such service attractive.

c) Power saving: To evaluate the power-saving characteristics of the PSPA/E protocol, some new variables should be taken into account: the acknowledgment duration E and the transmit-to-receive power ratio T/R of the terminals. We should note that, although power saving is the key issue, the PSPA/E protocol forces mobile terminals to transmit more frequently on the uplink channel in order to declare their reception availability. These extra transmissions may prove destructive unless low power transmitters are employed or, more correctly, unless the T/R factor is kept low. However, as we will see later, for all the practical T/R ratios and for reasonable values of E, the protocol always exhibits power-saving characteristics.

Curves in Fig. 12 represent the *normalized* downlink power consumption (NDPC) for the PSPA/E protocol. The NDPC is defined as the average power that a terminal needs to sustain communication on the downlink channel (i.e., to receive data packets from the base station) under the PSPA/E protocol, over the same power needed under the conventional M/M/1 system. The latter is equal to the receiver's power consumption (assumed to be one), since only the receiver is associated with the packet reception. On the other hand, in the PSPA/E protocol both the transmitter and the receiver are engaged to the packet reception; thus, the term "downlink power consumption" is considered more proficient than the "receiver power consumption."

In Fig. 12 the NDPC is shown versus the total downlink traffic demand for three different values of T/R power ratio and for two systems with different number of terminals ($N = 10, N = 20$). The acknowledgment transmission duration has chosen to be 0.02 and receiver's duty cycle is 0.1.

As we can see, the NDPC depends heavily on the traffic per user as well as on the T/R ratio. When the traffic per user increases, a terminal sends more acknowledgments and remains for longer periods in receive mode; therefore, its power consumption is increased too. Observe that the power consumption ramps up faster when the T/R ratio is large because more power is needed to accommodate acknowledgment transmissions. However, even when T/R is as much as 200 and the population is relatively low ($N = 10$), the NDPC stays well below 0.4, so, in this case, the power saving is always greater than 60%. In the extreme case where the system has a very large population, the NDPC tends to be almost flat and the value of T/R does not make a great difference (since acknowledgment transmissions are rare).

It is interesting to note that, when the ratio T/R is large, the NDPC starts to decrease from a critical traffic point and beyond. This is because the frequency of acknowledgment transmissions starts also to decrease and, at every service period, a terminal starts to receive multiple packets. (This is another advantage of exhaustive service related to power consumption.) Moreover, this critical traffic point becomes smaller as the T/R becomes larger (see Fig. 12). This beneficial property shows that the power-saving protocol tends to offset the destructive effects of high transmission power and to prevent excessive power consumption.

As we stated earlier, one advantage of the PSPA protocol is that terminals may control their power consumption according to their needs. This can be done either by requesting higher service priority or by reducing their duty cycle. These disciplines are examined in the following two cases.

d) Priority service: Now, imagine that every terminal is assigned a unique priority. This can be easily implemented in practice. For instance, when a new terminal registers with the base station, it could be assigned a priority value according to its current power resources. In this way the terminal with the lowest power resources could have the highest priority.

Priority is put into effect when two or more terminals acknowledge the reception of the same paging message. Then, instead of servicing these terminals randomly, the base station services them according to their priority. This means that high-priority terminals will be served sooner; therefore, their power consumption will be reduced comparative to the low-priority terminals. A similar observation may be stated for the mean packet delay.

In Table I we show the average number of terminals served each time a service period occurs, versus the total offered traffic λ. We observe that when the duty cycle q is small, then, on the average, we approximately serve one terminal at a time

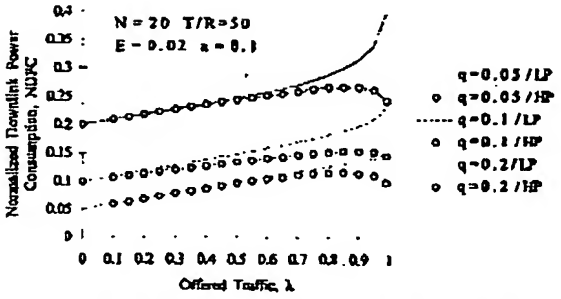

Fig. 13. Normalized downlink power consumption for various duty cycles and various priority levels. LP: lowest priority user, HP: highest priority user.

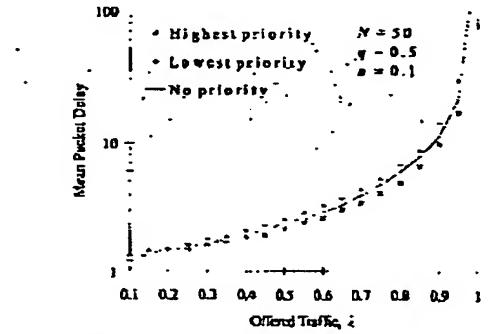

Fig. 14. Mean packet delay of the PSPA/E protocol for the highest and lowest priority users.

(at least under normal traffic condition). Only when q becomes relatively high and the population of the system N is large, will the priority service discipline have notable effects.

Fig. 13 demonstrates how priority affects the NDPC in a system with 20 users, each one having his own unique priority. We can observe that the NDPC of the highest priority user features a negative inclination at high loads, while the NDPC of the lowest priority user features a fast ramp up. This means that low-priority terminals may exhibit high power consumption at high traffic rates, especially when the T/R ratio becomes large. Also, as the system's population gets large, the low-priority terminals will consume more and more power, since they will be forced to cooperate with more and more terminals with higher priority. In fact, their NDPC will approximate unity because they will tend to remain constantly in receive mode when the downlink traffic gets high. This should be taken into account when scheduling to implement a priority service discipline with the PSPA/E protocol.

Fig. 14 shows the mean packet delay of the highest and lowest priority terminals in a network with 50 mobile terminals (again each terminal has a unique priority) that operate with duty cycle $q = 0.5$. Also, the curve that corresponds to random service discipline is illustrated, which lies in the middle of the two marginal priority cases.

At low traffic rates priority makes no difference since simultaneous acknowledgments rarely occur, but at high rates

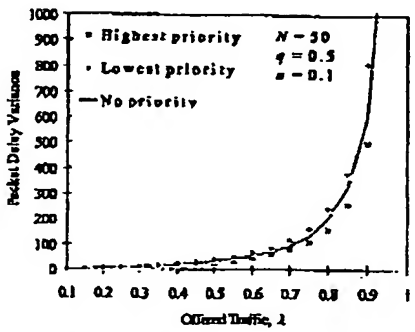

Fig. 15. Packet delay variance of the PSPA/E protocol for the highest and lowest priority users.

the high-priority users gain a notable advantage. Specifically, in the case shown in Fig. 14, the highest priority user features about 10%–20% smaller mean packet delay over the lowest priority user and about 8%–15% smaller mean packet delay over the random service discipline.

Similar statements hold true for the packet delay variance shown in Fig. 15. Since high-priority terminals are always served first, they will feature reduced packet delay variance relatively to random service. On the other hand, since low-priority terminals are always served at the end of every service period, they will be forced to wait for an additional random period, thus, they will feature increased packet delay variance. According to Fig. 15, the variance of the highest priority user may be reduced by almost 10%–20% (comparative to random service), while the variance of the lowest priority user may develop an increase of about 20%–30%.

e) Varying duty cycles: Now we are interested in studying how terminals with different duty cycles affect each other. Suppose there are two groups of terminals, the first operating with a duty cycle q_l and the second with a duty cycle q_h, and let $q_h > q_l$ (so, the terminals of the first group will consume less power). In such case, whenever many terminals are being paged, the terminals with low duty cycle q_l will have a performance disadvantage because, with a relatively high probability, a terminal with higher duty cycle q_h will acknowledge before them (since its sleep period will be smaller). Therefore, the group with duty cycle q_l will face longer paging delays relative to the other group and relative to the case where all terminals were operating with the same duty cycle q_l.

An interesting observation though is that terminals with different duty cycles will not heavily affect each other provided that their duty cycles reside in the flat range of Fig. 10 (say, larger than $q = 0.3$). In this range the sleep periods of terminals are not greatly different, so their degradation factors are almost identical. This very important result suggests that as long as mobile terminals operate with duty cycles larger than 0.3, their mutual affection is minimal and we can ignore it. In this case each terminal will face a performance almost identical to the performance where all the other terminals were using the same duty cycle with it. That result has been verified by

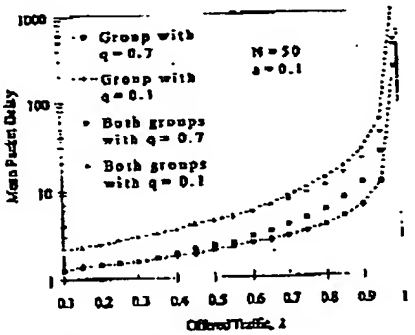

Fig. 16. Mean packet delay for two groups that implement different duty cycles.

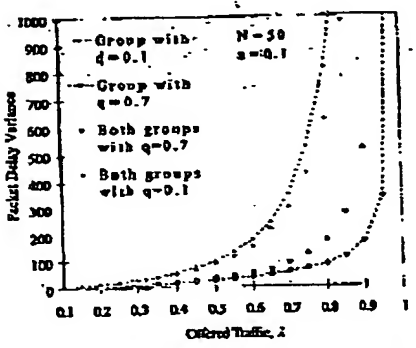

Fig. 17. Packet delay variance for two groups that implement different duty cycles.

our simulation model. Specifically, when q_0, q_1 were selected both larger than 0.3, there were no significant variations.

On the other hand, when some terminals utilize a quite small duty cycle, a strong affection with the other terminals is observed. This is illustrated in Fig. 16, where we consider a system with 50 mobile terminals split into two groups that implement different duty cycles, namely, 0.1 and 0.7. In this figure we see the mean packet delay of each group compared to the mean packet delay that we have considered so far (where all the terminal implement equal duty cycles). Note that the group with $q = 0.1$ develops an increased packet delay relative to the case where *all* terminals were operating with $q = 0.1$. On the other hand, the group with $q = 0.7$ demonstrates a substantial advantage since its mean packet delay is now reduced.

We also note that during low traffic conditions there is no significant effect (because approximately one terminal in being paged at a time), so in low and medium traffic networks the varying duty cycles are not important. However, as traffic becomes high, more and more terminals are simultaneously paged, so the disadvantage of the group with longer sleep periods begins to show up. Similar observations hold true for the variance of the packet delay, as we can see in Fig. 17.

As far as power consumption is concerned, no significant effect due to varying duty cycles is observed. This is quite

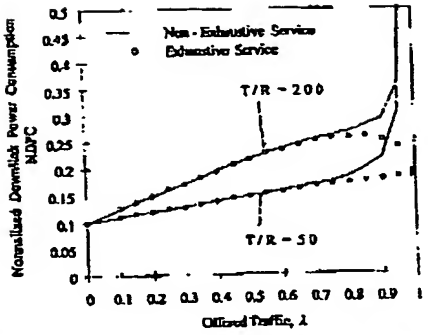

Fig. 18. Normalized downlink power consumption for the nonexhaustive service discipline compared to the exhaustive service discipline.

reasonable because having groups with different duty cycles introduce neither more acknowledgment transmissions nor longer receive periods, they just cause different paging delays.

Finally, the careful reader may note that a priority service discipline (as studied before) may mitigate the effects of different duty cycles since terminals with low duty cycles will have higher service priority over the terminals with high duty cycles.

2) Nonexhaustive Service: In order to increase the fairness of the in-band page-and-answer protocol, we consider the implementation of a nonexhaustive service discipline. The relevant flowchart is depicted in Fig. 7. This service mode is quite simplistic but in practice a much better scheme can be used to tradeoff throughput versus delay.

The main feature of this discipline is that at every service period *one* packet for every ready terminal[7] is transmitted. If there is residual buffered traffic for the ready terminals, another service period will follow *after* the transmission of a paging message. So, a terminal with high buffered traffic will not monopolize the downlink channel for a long time period as in the exhaustive service discipline. In fact, the base station will periodically suspend its service and will transmit a paging message to keep the rest of terminals informed about the total downlink demand. Therefore, other terminals will also have the chance to get ready and to share the downlink channel.

However, in the nonexhaustive service discipline a terminal will generally remain for longer periods in receive mode, so a power consumption cost is unavoidable. This is shown in Fig. 18, where we see that, at high traffic rates, mobile terminals consume more power relatively to the exhaustive service. In fact, the NDPC in the nonexhaustive service approximates unity as traffic gets high (since terminals tend to remain continuously in operation).

We are in the stage of setting up an experimental network that will operate with the in-band page-and-answer protocol. We have already built the proper hardware [19] and we are now programming the protocol itself. Various enhancements

[7] A ready terminal is a terminal that it is known to be in receive mode and awaiting its pending traffic. A ready terminal goes back into power-saving mode only after having no residual buffered traffic.

are scheduled in order to make the power-saving protocol more efficient and more robust.

Also, the power consumption of the terminals is expected to be smaller than we have shown because a terminal is expected to be less persistent on searching for paging messages. (In this paper we assumed that after waking up, a terminal *always* seeks for a paging message, and thus *always* consumes some energy for monitoring. However, this is quite pessimistic and costly.) We base this on mainly two reasons.

- First, there would be instances that a terminal wakes up and detects no downlink traffic. In such cases, it is better to immediately shut down, rather than keep on monitoring and waiting for a timeout, because the probability to arrive a message *for it* in the middle of its monitoring time is small.
- Second, if a terminal wakes up during a data transmission period, there is no reason to search for a paging message. Instead, it may power down and schedule a later wakeup. In order to act so, we need to provide some means to aid terminals in identifying if they are receiving data or paging. One way that we plan to accomplish this is to have the base station transmit a subtone together with the paging messages. So, if a terminal fails to detect the subtone, it deduces that it receives data. This is really a fast and simple method to distinguish between data and paging periods and it saves valuable power.

V. Conclusion

The page-and-answer protocol that we have considered in this paper features significant power-saving characteristics at a cost of increased mean packet delay and increased packet delay variance. We have shown that, if the duration of paging messages is kept small, the packet delay degradation and variance degradation are minimized. Also, the protocol exhibits a superior performance over time-division multiple access (TDMA)[8] (that inherently provides power-saving characteristics) as long as paging messages have reasonably small duration.

The most profound potentials of the protocol are established under low traffic conditions, where the mean traffic per user is small. At these traffic rates, both the mean packet delay and the packet delay variance feature low degradation, and also considerable power saving is achieved. For this reason, patch-type applications, like two-way messaging, small file transfer, and mail exchange, are likely to be the most suitable applications.

Moreover, a low transmitter to receiver power ratio is desired for mitigating the effects of acknowledgment transmissions to the power consumption characteristics. So, networks employing picocellular and microcellular arrangements like wireless local area networks (LAN's), are considered as the most favorable networks for implementation, because the transmission power in these systems is typically low, in the order of megawatts. There are also other reasons that render these systems suitable for the page-and-answer protocol—they feature an extremely small round trip delay; thus, they ensure a fast returning channel, which is critical for the acknowledgment process. Also, the negligible corruption probability assumed for the acknowledgment packets is realized more easily in these networks because they usually operate on high-capacity channels and, therefore, the transmission duration is minimal.

Along with further enhancements, the considered protocol can be envisaged as an optional power-saving feature, where terminals that run out of energy switch to power saving after notifying the central base station. The protocol provides the means for independent power management facilities, which are of considerable importance in a personal communications environment. In such an environment every mobile terminal will be able to retain a dynamically tunable duty cycle, proportional to its current battery status, and apply a distinct network-independent algorithm to manage its own power resources. In this case (as we have shown) a terminal will only be slightly affected by other terminals that implement different duty cycles provided that the duty cycles are not very small. Finally, if the network allows priority scheduling, then priority can also be used to facilitate the power saving/performance options of a terminal.

[8] A direct comparison with TDMA is not desirable since TDMA is known to be inefficient for data communication.

References

[1] A. S. Hoa, "An introduction to paging—What it is and how it works," *Motorola Electronics Pte. Ltd.*, 1994.

[2] *MOBITEX Interface Specification*, RAM Mobile Data, 1993.

[3] A. K. Salkintzis and C. Chamzas "Mobile packet data technology: A survey of MOBITEX," *IEEE Personal Commun.*, vol. 4, pp. 10–18, Feb. 1997.

[4] J. Mallo Jr. and P. Wayner, "Wireless mobile communication," *Byte*, pp. 147–154, Feb. 1993.

[5] H. Kobayashi and A. G. Konheim, "Queueing models for computer communications system analysis," *IEEE Trans. Commun.*, vol. COM-25, pp. 2–28, 1977.

[6] N. Abramson, "Multiple access in wireless digital networks," *Proc. IEEE*, vol. 82, pp. 1360–1369, Sept. 1994.

[7] S. S. Lam, "Delay analysis of a time division multiple access (TDMA) channel," *IEEE Trans. Commun.*, vol. COM-25, pp. 1489–1494, Dec. 1977.

[8] J. E. Padgett, C. G. Guenther, and T. Hattori, "Overview of wireless personal communications," *IEEE Commun. Mag.*, vol. 33, pp. 28–41, Jan. 1995.

[9] L. Kleinrock, *Queueing Systems Vol. I: Theory* New York: Wiley-Interscience, 1975.

[10] ———, *Queueing Systems Vol. II: Computer Applications*. New York: Wiley-Interscience, 1976.

[11] R. Steele, *Mobile Radio Communications*. Piscataway, NJ: IEEE Press–Pentech Press, 1994.

[12] *CDPD Interface Specification*, CDPD Forum, Jan. 1995.

[13] *Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Draft Standard 802.11 D3.1, Apr. 1996.

[14] D. Makishima and B. Obara, "IEEE 802.11 standards create new options for wireless LAN connectivity," *Mobile Comput. Technol.*, p. 23–37, Nov./Dec. 1995.

[15] S. Chatzis, A. K. Salkintzis, and C. Chamzas, "An energy saving protocol for data networks using out-of-band signalling," in *5th Pan-Hellenic Computer Science Conf.*, Athens, Greece, Dec. 7–9, 1995

[16] A. K. Salkintzis, C. Chamzas, and C. Koukourlis, "An energy saving protocol for mobile data networks," in *Int. Conf. Advances in Communication and Control (COMCON 5)*, June 26–30, 1995.

[17] A. K. Salkintzis, S. Chatzis, and C. Chamzas, "An energy-efficient protocol for mobile computing environments," in *Int. Workshop on Mobile Communications*, Thessaloniki, Greece, Sept. 19–21, 1996.

[18] A. K. Salkintzis and C. Chamzas, "An outband paging protocol for energy-efficient mobile computing," *IEEE Trans. Commun.*, submitted for publication.

[19] A. K. Salkintzis, J. Plevridis, C. Koukourlis, and C. Chamzas, "Design and implementation of a low-cost wireless network for remote control and monitoring applications," *Microprocessors Microsyst.*, vol. 21, no. 2, pp. 79–88, Oct. 1997.
[20] R. E. Kahn et al., "Advances in packet radio technology," *Proc. IEEE*, vol. 66, pp. 1468–1496, Nov. 1978.
[21] F. A. Tobagi *et al.*, "Modeling and measurement techniques in packet communication networks," *Proc. IEEE*, vol. 66, pp. 1423–1447, Nov. 1978.
[22] K. Pahlavan and A. H. Levesque, "Wireless data communications," *Proc. IEEE*, vol. 82, pp. 1398–1430, Sept. 1994.

Apostolis K. Salkintzis was born in Heraklion, Greece. He received the Diploma degree in electrical engineering in 1991 and the Ph.D. degree in 1997, both from the Electrical and Computer Engineering Department, Democritus University of Thrace, Xanthi, Greece.

Since 1992 he has been working on various research projects with the Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, Greece. His primary research interests are in the areas of digital communication systems and PCS. Within the area of digital communications, he is interested in error-correction coding, adaptive channel equalization, and mobile channel modeling, as well as designing and implementing efficient and low-cost radio modems. Within the PCS area, he is interested in mobility management, access signaling procedures, and efficient air-interface protocols.

Dr. Salkintzis is a member of the Technical Chamber of Greece.

Christodoulos Chamzas (S'75–M'79–SM'85) was born in Komotini, Greece. He received the Diploma Degree in electrical and mechanical engineering from the National Technical University of Athens, Athens, Greece, in 1974, and the M.S. and Ph.D. degrees in electrical engineering from the Polytechnic Institute of New York, Farmingdale, in 1975 and 1979, respectively.

From 1979 to 1982 he was an Assistant Professor with the Department of Electrical Engineering, Polytechnic Institute of New York, Farmingdale. In September 1982 he joined AT&T Bell Laboratories, Holmdel, NJ, where he was a member of the Visual Communications Research Department until 1990. Since September 1990 he has been a member of the Faculty of the Electrical Engineering Department, Democritus University of Thrace, Xanthi, Greece, where he is a Director of the Electric Circuits Analysis Laboratory. He has been a major player in the definition, design, and implementation of the CCITT/ISO (JBIG, JPEG, etc.) standards for coding, storage, and retrieval of images (color and bilevel), an area in which he holds six international patents. In 1985–86 he was a Visiting Professor with the Department of Computer Science, University of Crete, Heraklion, Greece. His primary interests are in digital signal processing, image coding multimedia, and communications systems. He is an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS.

Dr. Chamzas is a member of the Technical Chamber of Greece and Sigma Xi.